Collusion-resistant, Incentive-compatible Feedback Payments

Radu Jurca
Ecole Polytechnique Fédérale de Lausanne
(EPFL)
Artificial Intelligence Laboratory
CH-1015 Lausanne, Switzerland
radu.jurca@epfl.ch

Boi Faltings
Ecole Polytechnique Fédérale de Lausanne
(EPFL)
Artificial Intelligence Laboratory
CH-1015 Lausanne, Switzerland
boi.faltings@epfl.ch

ABSTRACT

Online reputation mechanisms need honest feedback to function effectively. Self-interested agents report the truth only when explicit rewards offset the potential gains obtained from lying. Feedback payment schemes (monetary rewards for submitted feedback) can make truth-telling rational based on the correlation between the reports of different buyers.

In this paper we investigate incentive-compatible payment mechanisms that are also resistant to collusion: groups of agents cannot collude on a lying strategy without suffering monetary losses. We analyze several scenarios, where, for example, some or all of the agents collude. For each scenario we investigate both existential and implementation problems. Throughout the paper we use *automated mechanism design* to compute the *best* possible mechanism for a given setting.

Categories and Subject Descriptors

I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence

General Terms

Algorithms, Design, Economics

Keywords

reputation mechanisms, mechanism design, incentive compatibility, collusion resistance

1. INTRODUCTION

Users increasingly resort to online feedback forums (reputation mechanisms) for obtaining information about the products or services they intend to purchase. The testimonies of previous buyers disclose hidden, *experience-related* [19], product attributes (e.g., quality, reliability, ease of use, etc.) that can only be observed after the purchase. This previously unavailable information allows the buyers to make better, more efficient decisions.

A key ingredient for all reputation mechanisms is honest feedback. Human users exhibit high levels of altruistic (i.e., honest) reporting, despite empirical evidence that lying can bring external benefits [8, 21]. Nevertheless, the future online economy may be dominated by rational, utility maximizing software agents, that will exploit such misreporting opportunities. Hence the need for designing reputation mechanisms that are incentive-compatible: i.e., rational agents find it in their best interest to report the truth.

Fundamental results in the mechanism design literature [7, 5] show that side payments can be designed to create the incentive for agents to report their private opinions truthfully. The best such payment schemes have been constructed based on *proper scoring rules* [15, 12, 2], and exploit the correlation between the observations of different buyers about the same good.

Miller, Resnick and Zeckhauser [18] adapt these results to online feedback forums. A central processing facility (the reputation mechanism) *scores* every submitted feedback by comparing it with another report (called the *reference* report) about the same good. They prove the existence of general incentive-compatible payment mechanisms that create an equilibrium where the return when reporting honestly is better by at least an arbitrary margin δ.

Jurca and Faltings [14] use an identical setting to apply *automated mechanism design* [3, 20]. Incentive-compatible payments are computed by solving an optimization problem with the objective of minimizing the required budget. The simplicity of specifying payments through closed-form scoring rules is sacrificed for significant gains in efficiency.

Intuitively, payment mechanisms encourage truth-telling because reporters expect to get paid according to how well their feedback improves the current predictor of the reference report. Every feedback report modifies the reputation information, which acts as a predictor for future observations. The payment received by the reporter reflects the quality of the updated predictor, tested against the reference report. Assuming that the reference report is honest, every agent has the incentive to update the current reputation such that it mirrors her subjective beliefs. Agents thus report honestly, and truth-telling is a Nash equilibrium.

For example, consider the owner (she) of a new house who needs some plumbing work done. There are good and bad plumbers; the one chosen by our owner has a fairly good

EC'07, June 11–15, 2007, San Diego, California, USA.
Copyright 2007 ACM 978-1-59593-653-0/07/0006 ...$5.00.

reputation that predicts high quality work with probability 75%. Once the work gets done, the owner will have a new (private) belief regarding the reputation of the plumber. If she is happy with the service, she believes the plumber should have an even better reputation (that predicts, for example, good service with probability 87%[1]). On the other hand, if the owner is dissatisfied with the service, she believes that the plumber should have a lower reputation (the probability of good service should only be 39%). An online reputation mechanism asks the owner to share feedback about the plumber, and proposes the following payment rule: "*the submitted report is paid only if it matches the report submitted by another client about the same plumber. A negative report will be paid* \$2.62*, while a positive report will be paid* \$1.54". One can verify that honest reporting maximizes the expected payment of the owner, regardless of the actual experience[2].

While honest reporting is a Nash Equilibrium (NEQ), so is always reporting negative (or positive) feedback. Moreover, the expected payoff from these constant reporting strategies (\$2.62 or \$1.54 respectively) are both higher than the expected payoff from the honest equilibrium. Unfortunately, the existence of multiple equilibria is not an isolated problem specific to our example. In previous work [13] we show that all binary incentive-compatible payment mechanisms suffer from the same drawback: there are lying equilibria that generate higher expected payoffs than the truthful equilibrium.

This brings forth the problem of collusion. Rational agents have no reason to report truthfully, if they can do better by coordinating on a lying equilibrium with higher payoff. Hence the motivation of this paper: we investigate incentive-compatible payment mechanisms that are also resistant to collusion. The extent and the power of the coalition influences the complexity and difficulty of the design problem.

We will consider four collusion scenarios: First, we consider complete coalitions (all agents may be part of the coalition) where agents may not redistribute revenues, and may only collude on symmetric strategies (all colluders report according to the same strategy). We obtain a positive result and show that by using several reference reports, it is possible to construct a payment mechanism with a unique honest reporting symmetric equilibrium.

Second, we consider a close to worst case scenario, where all agents may collude, and coordinate on different reporting strategies (every agent may report according to a different strategy). Unsurprisingly, the result we obtain here is negative: regardless of the number of reference reports, no incentive-compatible payment mechanism has a unique honest reporting equilibrium.

Third, we move to more realistic scenarios where only a fraction of the agents may collude. Colluders may not transfer money among themselves, but may fully coordinate their reporting strategies. We show that honest reporting can be made a dominant strategy when (a) the coalition size is small enough, and (b) the non-colluders are reporting honestly.

Finally, we consider what happens when a single strategic entity controls a number of fake online identities. As colluders can now transfer money among themselves, the payment mechanism must ensure that the cumulative revenue of the coalition is maximized by the honest reports. The result is positive, and given that non-colluders report honestly, appropriate payments elicit truthful information from the coalition as a whole.

An important ingredient of our solution is to use several *reference reports* when computing the feedback payments. This not only decreases the total cost of incentive-compatibility (a contribution of our previous work [14]), but also allows the design of mechanisms where honest reporting is the unique equilibrium.

After mentioning related work, this paper proceeds as follows. Section 2 formally introduces our model, Section 3 introduces incentive-compatible payment mechanisms and presents some of their properties. Sections 4 through 7 each treat one collusion scenario. Finally we discuss future work and conclude.

1.1 Related Work

Our work relates to the literature on (computational) mechanism design, implementation theory, and incentive contracts for principal-(multi)agent settings. The literature on mechanism design (see [11] for a survey) and implementation theory (see [10] for a survey) addresses the design of mechanisms and institutions that satisfy certain properties, given that the agents using the mechanism behave strategically. The main difference between mechanism design and implementation theory is that of multiple equilibria. In *mechanism design* literature, the goal of the designer is to find the mechanism that has the desired outcome as *an* equilibrium. In the *implementation* literature, on the other hand, the mechanism is required to have *only* the desired equilibria. From this perspective, our results are closer to the implementation theory, since in our quest for collusion resistance we look for payment mechanisms that induce honest reporting as the only (or in some sense the best) equilibrium.

Computational mechanism design [6] extends the classical literature by looking for equilibria that are also computationally attractive. In the feedback reporting setting, however, the main computational problem resides in computing the mechanism itself (i.e., the payments), and not in executing it. In this paper, we rely on *automated mechanism design* [3] to compute the best mechanism for each context. The objective of the designer is to minimize the budget required to pay for feedback, while enforcing the incentive-compatibility constraints. The design problem is a linear optimization problem and can be solved efficiently [14]. A related method is *incremental mechanism design* [4] where the mechanism is solved iteratively, by incrementally adding new constraints.

A number of papers discuss incentive contracts that a principal should offer to several agents whose effort levels are

[1] The rationale behind these numbers will become apparent in Section 2

[2] If the owner experiences good service from the plumber, she expects that some other client also gets good service with probability 87%. Assuming that the other client reports truthfully, the owner's expected payment is: $.87 \cdot 1.54 + .13 \cdot 0 = 1.34$ if she reports good service, or $.87 \cdot 0 + .13 \cdot 2.62 = 0.34$ if she reports bad service; Likewise, if the owner experiences bad service, she expects that the reference report will be negative with probability $1 - .39 = .61$. In this case, her expected payment is: $.39 \cdot 1.54 + .61 \cdot 0 = 0.6$ if she reports good service, or $.39 \cdot 0 + .61 \cdot 2.62 = 1.6$ if she reports bad service. In both cases, honest reporting is better than lying by \$1.

private. The reward received by each agent depends on the output observed by the principal, and on the declarations of other agents. [9], [17], and [16] show that efficient contracts exist that are also incentive-compatible and collusion-proof. While the feedback reporting problem is similar, it differs in one major aspect: the mechanism designer (i.e., the principal) does not observe a direct signal which is correlated to the reporters' (i.e., agents') private information.

2. THE MODEL

We consider an online market where a number of rational buyers (or agents) experience the same product (or service). The quality of the product remains fixed, and defines the product's (unknown) *type*. Θ is the finite set of possible types, and θ denotes a member of this set. We assume that all buyers share a common belief regarding the prior probability $Pr[\theta]$, that the product is of type θ. $\sum_{\theta \in \Theta} Pr[\theta] = 1$.

After the purchase, every buyer perceives a binary signal about the quality (i.e., true type) of the product. Quality signals are denoted as 1 (high quality) and 0 (low quality), meaning that the buyer was satisfied, respectively dissatisfied with the product. Every product type is characterized by a different probability distribution over the signals perceived by the buyers. Let $Pr[1|\theta]$ be the probability that the agent buying a product of type θ is satisfied (i.e., observes the quality signal 1). $Pr[1|\theta_1] \neq Pr[1|\theta_2]$ for all $\theta_1 \neq \theta_2 \in \Theta$, and $Pr[1|\cdot]$ is assumed common knowledge.

A central reputation mechanism asks every buyer to submit feedback. Buyers are assumed rational, and not constrained to report the truth. The set of pure reporting strategies of a buyer is $A = \{(a_0, a_1) | a_0, a_1 \in \{0, 1\}\}$, where $a = (a_0, a_1)$ denotes the strategy according to which the buyer announces $a_0 \in \{0, 1\}$ when she observes low quality, and $a_1 \in \{0, 1\}$ when she observes high quality. We will often call the reports 0 and 1 as the *negative*, respectively the *positive* report.

To ease the notation, we name the four members of the set A as the $h(onest)$, $l(ie)$, $all1$ and $all0$ strategies:

- $h = (0, 1)$ is the honest reporting strategy;
- $l = (1, 0)$ is the strategy of always lying: the buyer reports 1 instead of 0 and 0 instead of 1;
- $all1 = (1, 1)$ is the strategy of always reporting 1;
- $all0 = (0, 0)$ is the strategy of always reporting 0;

The reputation mechanism pays buyers for the submitted reports. The amount received by buyer i can depend on any information available to the reputation mechanism: namely, the reports submitted by other buyers, and the common knowledge regarding the environment (probability distribution over types, and conditional probability distributions of quality signals). Let N be the total number of reports available to the reputation mechanism. N is finite, either because the total number of buyers is finite, or because the reputation mechanism cannot wait indefinitely to receive more reports.

Note that the reputation mechanism (a) does not know the true type of the product, and (b) cannot purchase the product in order to get some first-hand experience regarding its quality.

Discarding from the notation the dependence on the common knowledge, a payment mechanism (employed by the reputation mechanism) is a function $\tau : \{0, 1\} \times \{0, 1\}^{N-1} \to \mathbb{R}^+$, where $\tau(\alpha_i, \alpha_{-i}) \geq 0$ is the amount paid to buyer i when she reports $\alpha_i \in \{0, 1\}$ and the other $N - 1$ buyers report $\alpha_{-i} \in \{0, 1\}^{N-1}$. The reports α_{-i} are also called the *reference reports* of agent i, since they constitute the reference for computing the payment for agent i. Payments are non-negative because most online forums do not have the means to impose punishments on the reporters.

As the order of reports is not important, we can simplify the payment mechanism by assuming that $\tau(\alpha_i, \alpha_{-i}) = \tau(\alpha_i, \alpha^*_{-i})$ for all α_{-i} and α^*_{-i} that contain the same number of positive reports. A more compact description of the payment mechanism is thus given by the amounts $\tau(\alpha, n)$ where $n \in \{0, 1, \ldots, N - 1\}$ is the number of positive reports submitted by the reference reporters.

The payoff expected by agent i depends on the distribution of the reference reports. If the other agents report honestly, the distribution of the reference reports can be computed from the prior beliefs, and the true observation, $o_i \in \{0, 1\}$ of agent i. The probability that exactly n positive reports were submitted by the other $N - 1$ agents is:

$$Pr[n|o_i] = \sum_{\theta \in \Theta} Pr[n|\theta] Pr[\theta|o_i]; \qquad (1)$$

where $Pr[n|\theta]$ is given by the binomial distribution, and $Pr[\theta|o_i]$ can be computed from Bayes' Law:

$$Pr[n|\theta] = \binom{N-1}{n} Pr[1|\theta]^n \big(1 - Pr[1|\theta]\big)^{N-1-n};$$

$$P[\theta|o_i] = \frac{Pr[o_i|\theta] Pr[\theta]}{Pr[o_i]}; \quad Pr[o_i] = \sum_{\theta \in \Theta} Pr[o_i|\theta] Pr[\theta];$$

A strategy profile a is a vector $(a_i)_{i=1,\ldots,N}$, prescribing the reporting strategy $a_i \in A$ for each agent i. We will sometimes use the notation $a = (a_i, a_{-i})$, where a_{-i} is the strategy profile for all agents except i; i.e., $a_{-i} = (a_j)$, for $j = 1, \ldots, i - 1, i + 1, \ldots, N$. Given the profile of reporting strategies (a_i, a_{-i}), let $\pi[n, a_{-i}]$ describe the belief of agent i regarding the distribution of the reference reports, when:

- n out of the other $N-1$ agents observe the high quality signal, 1
- the other $N - 1$ agents are reporting according to the strategy profile a_{-i};

Given n and a_{-i}, agent i believes with probability $\pi[n, a_{-i}](x)$ that x reference reports are positive. If $a_i(o_i) \in \{0, 1\}$ is the value of the report prescribed by strategy a_i given the true observation o_i, the expected payoff to agent i is:

$$V(a_i, a_{-i}|o_i) = \sum_{n=0}^{N-1} Pr[n|o_i] \sum_{x=0}^{N-1} \pi[n, a_{-i}](x) \tau(a_i(o_i), x); \quad (2)$$

Before moving to the next section, we can now justify the numerical values chosen for the example presented in the introduction. The two possible types of the plumber are *good* (θ_G) and *bad* (θ_B), the good type being more likely than the bad type: $Pr[\theta_G] = 0.8$ and $Pr[\theta_B] = 0.2$. The good plumber provides good service with high probability $Pr[1|\theta_G] = 0.9$; the bad plumber provides good service with the lower probability, $Pr[1|\theta_B] = 0.15$. The prior *reputation* of the plumber predicts a good service with probability: $Pr[\theta_G]Pr[1|\theta_G] + Pr[\theta_B]Pr[1|\theta_B] = 0.75$. However, the posterior reputation depends on the agent's actual experience.

If the agent observes 1, the posterior belief regarding the type of the plumber will be: $Pr[\theta_G|1] = 1 - Pr[\theta_B|1] = 0.96$ (computed by Bayes' Law), and the probability that the plumber provides good service to another client is: $Pr[1|1] = Pr[1|\theta_G]Pr[\theta_G|1] + Pr[1|\theta_B]Pr[\theta_B|1] = 0.87$. Likewise, if the agent observes 0, the posterior belief regarding the type of the plumber is: $Pr[\theta_G|0] = 1 - Pr[\theta_B|0] = 0.32$, and the probability that the plumber provides good service to another client is: $Pr[1|0] = Pr[1|\theta_G]Pr[\theta_G|0] + Pr[1|\theta_B]Pr[\theta_B|0] = 0.39$.

3. INCENTIVE-COMPATIBLE PAYMENT MECHANISMS

A payment mechanism is *incentive-compatible* when honest reporting is a Nash Equilibrium (NEQ): i.e., no agent can gain by lying when other agents report honestly. Formally, let (h_i, h_{-i}) be the strategy profile where all agents report honestly. It is optimal for agent i to report the truth if and only if, for any observation o_i, the honest report maximizes the agent's expected payoff:

$$V(h_i, h_{-i}|o_i) > V(a_i, h_{-i}|o_i)$$

for any reporting strategy $a_i \in A\backslash\{h\}$, and any observation $o_i \in \{0,1\}$.

Since reference reports are truthful, the expected payoff to agent i is:

$$V(h_i, h_{-i}|o_i) = \sum_{n=0}^{N-1} Pr[n|o_i]\tau(o_i, n);$$

and the incentive-compatibility constraints become:

$$\sum_{n=0}^{N-1} Pr[n|o_i]\tau(o_i, n) > \sum_{n=0}^{N-1} Pr[n|o_i]\tau(1 - o_i, n); \qquad (3)$$

for $o_i = 0, 1$.

Practical mechanisms require certain margins for truth-telling [14]. Honest reporting must be better than lying by at least some margin δ, chosen by the mechanism designer to offset the external benefits an agent might obtain by lying. Rewriting (3) to account for the margin δ, an incentive-compatible payment mechanism satisfies the constraints:

$$\begin{aligned} &\sum_{n=0}^{N-1} Pr[n|1]\Big(\tau(1,n) - \tau(0,n)\Big) \geq \delta; \\ &\sum_{n=0}^{N-1} Pr[n|0]\Big(\tau(0,n) - \tau(1,n)\Big) \geq \delta; \end{aligned} \qquad (4)$$

formalizing the intuition that it is more profitable to report positively (respectively negatively) when observing high (respectively low) quality.

[15], and [18] show that it is possible to construct payment mechanisms that satisfy the constraints in (4), based on *scoring rules*. Jurca and Faltings [14] build on this existence result and describe an algorithm that computes the optimal (i.e., budget minimizing) payment mechanism. We will use this latter approach in this paper, for the obvious practical advantages of designing an incentive compatible reputation mechanism as cheaply as possible.

The expected payment to an honest reporter (in the truthful NEQ) is the weighted sum between the expected payment to an agent that truthfully reports 1, and the expected payment to an agent that truthfully reports 0:

$$W = Pr[1]\sum_{n=0}^{N-1} Pr[n|1]\tau(1,n) + Pr[0]\sum_{n=0}^{N-1} Pr[n|0]\tau(0,n); \quad (5)$$

where $Pr[1]$ (respectively $Pr[0]$) are the prior probabilities that the agent will perceive high (respectively low) quality, and are defined as: $Pr[o_i] = \sum_{\theta\in\Theta} Pr[o_i|\theta]Pr[\theta]$.

The payment scheme that minimizes the budget required to pay for one honest report therefore solves the linear optimization problem:

LP 1.

$$\begin{aligned} min \quad & W = Pr[1]\sum_{n=0}^{N-1} Pr[n|1]\tau(1,n) + Pr[0]\sum_{n=0}^{N-1} Pr[n|0]\tau(0,n); \\ s.t. \quad & \sum_{n=0}^{N-1} Pr[n|1]\Big(\tau(1,n) - \tau(0,n)\Big) \geq \delta; \\ & \sum_{n=0}^{N-1} Pr[n|0]\Big(\tau(0,n) - \tau(1,n)\Big) \geq \delta; \\ & \tau(0,n), \tau(1,n) \geq 0; \forall n = \{0, 1, \ldots, N-1\}; \end{aligned}$$

Although numerical algorithms can efficiently solve LP 1, the analytical solution helps us gain additional insights about the structure of incentive-compatible payment mechanisms. It turns out that LP 1 has a simple solution (details in Appendix A) where:

$$\begin{aligned} &\tau(0,n) = 0, \forall n \neq n_1; \quad \tau(1,n) = 0, \forall n \neq n_2 \\ &\tau(0,n_1) = \delta\frac{Pr[n_2|0] + Pr[n_2|1]}{Pr[n_2|1]Pr[n_1|0] - Pr[n_2|0]Pr[n_1|1]}; \\ &\tau(1,n_2) = \delta\frac{Pr[n_1|0] + Pr[n_1|1]}{Pr[n_2|1]Pr[n_1|0] - Pr[n_2|0]Pr[n_1|1]}; \\ &n_1 = \arg\min_n \frac{Pr[n|1]}{Pr[n|0]}; \quad n_2 = \arg\min_n \frac{Pr[n|0]}{Pr[n|1]} \end{aligned}$$

Intuitively, the optimal payment mechanism does not pay the negative or positive report of an agent unless the reference reports contain exactly n_1, respectively n_2 positive reports. The values n_1 and n_2 are chosen such that the posterior belief of the reporter regarding the reference reports changes the most:

- $Pr[n_1|0]$ increases the most with respect to $Pr[n_1|1]$: e.g., $n_1 = \arg\min_n Pr[n|1]/Pr[n|0]$;
- $Pr[n_2|1]$ increases the most with respect to $Pr[n_2|0]$: e.g., $n_2 = \arg\min_n Pr[n|0]/Pr[n|1]$;

The values of $\tau(0,n_1)$ and $\tau(1,n_2)$ are then computed to guarantee the margin δ for honest reporting.

A similar property holds for all payment mechanisms[3] that satisfy the incentive compatibility constraints: there must be at least two values of the reference reports, $n_1 \neq n_2$, such that:

$$\begin{aligned} \tau(0,n_1) > \tau(1,n_2), \quad Pr[n_1|0] > Pr[n_1|1], \\ \tau(1,n_2) > \tau(0,n_2), \quad Pr[n_2|1] > Pr[n_2|0]; \end{aligned}$$

[3] One might wish, for example, to design a mechanism that minimizes the expected budget paid to all N buyers. In this case, the objective function of the problem LP 1 is: $\bar{W} = \sum_{n=0}^{N} Pr[n]\big(n \cdot \tau(1, n-1) + (N-n) \cdot \tau(0,n)\big)$, where $Pr[n]$ is the prior probability that n out of N buyers observe high quality;

When $\tau(0, n_1)$, $\tau(1, n_2)$, $\tau(1, n_1)$ and $\tau(0, n_2)$ are scaled appropriately, a rational agent prefers the 'bet' on n_1 when she observes low quality, and the 'bet' on n_2 when she observes high quality.

It is exactly this property that makes it impossible to design an incentive-compatible mechanism that has honest reporting as the unique (or the most preferred) NEQ with only one reference report. n_1 and n_2 are constrained to take the values 0, respectively 1, since by Bayes' Law, $Pr[0|0] > Pr[0|1]$ and $Pr[1|1] > Pr[1|0]$. This results in positive payments $\tau(0, 0) > \tau(0, 1)$ and $\tau(1, 1) > \tau(1, 0)$ (as pointed out in the example from the introduction), thus the constant reporting strategies (always reporting 1 or always reporting 0) are also Nash Equilibria. Honest reporting is rewarded by a linear combination of $\tau(0, 0)$ and $\tau(1, 1)$, so at least one of the constant reporting strategies is more attractive than truth-telling. [13] formally develops this result.

Using several reference reports decreases the budget required to pay the reporters [14], and sometimes allows to design incentive-compatible payments where honesty is the only (or the most attractive) NEQ. In the following sections we explore some of the settings where such results apply.

4. NON-TRANSFERABLE UTILITIES, NO COORDINATION

The simplest collusion scenario (from the perspective of the mechanism designer) is to assume that agents (a) can only coordinate once (before any of them purchases the product) on the same (pure) reporting strategy, and (b) they cannot transfer payments from one another. Intuitively, this setting characterizes anonymous feedback forums where the colluders do not have side-channels for exchanging information. The absence of communication channels is not an underlying assumption about the physical world, but rather a contextual implication: most online buyers do not know who is going to buy the same product in the immediate future, and therefore cannot synchronize their reports.

Nevertheless, the agents collude in the sense that they *all* have one access to a trusted oracle that gives them a reporting strategy. For example, the role of the oracle might be played by a trustworthy site that analyzes the reporting strategies and recommends the best one.

The lack of coordination between colluders considerably simplifies the problem of the mechanism designer. The only supplementary constraint on the incentive-compatible payment mechanism is to ensure that none of the pure symmetric strategy profiles is a NEQ.

The set of pure strategies is finite (and contains 3 *lying* strategies) therefore we can exhaustively enumerate the constraints that prevent the corresponding symmetric lying strategy profiles to be NEQ. Since agents cannot transfer payments from one another, the constraints on the payments should simply provide incentives for deviating from the collusion strategy:

- $all1$ (always reporting 1) is not NEQ when a rational agent would rather report 0 instead of 1 given that all other agents follow $all1$:

$$\tau(0, N-1) > \tau(1, N-1); \qquad (6)$$

- $all0$ (always reporting 0) is not NEQ when a rational agent would rather report 1 instead of 0 given that all other agents follow $all0$;

$$\tau(1, 0) > \tau(0, 0); \qquad (7)$$

- $l(ie)$ is not NEQ when at least one agent (either observing 1 or 0) would rather report the truth. Given that other agents always lie, $N - 1 - n$ reference reports will be positive whenever n high quality signals were actually observed:

$$\begin{aligned} &\text{either} \sum_{n=0}^{N-1} Pr[n|0]\big(\tau(0, N-1-n) - \tau(1, N-1-n)\big) > 0; \\ &\text{or} \sum_{n=0}^{N-1} Pr[n|1]\big(\tau(1, N-1-n) - \tau(0, N-1-n)\big) > 0; \end{aligned} \qquad (8)$$

The objective function (5), and the constraints (4), (6), (7) and (8) define the optimal incentive-compatible payment mechanism that is also collusion-resistant in the sense explained in the beginning of this section (i.e., honest reporting is the unique pure-strategy symmetric NEQ). To compute the payments, the mechanism designer must solve two linear optimization problems, one corresponding to each branch of the constraint (8).

A collusion-resistant mechanism is easier to find when the number of reports available to the reputation mechanism is higher. The minimum number of reference reports that guarantee the existence of a collusion-proof payment mechanism depends on the distributions $Pr[n|\cdot]$, and on the context.

For the example provided in the introduction, it takes $N = 4$ reports to design a collusion-resistant payment mechanism. The expected distribution over reference reports is: $Pr[0\ldots3|0] = [0.4179, 0.2297, 0.1168, 0.2356]$ when the plumber provides bad service, and $Pr[0\ldots3|1] = [0.0255, 0.0389, 0.2356, 0.7]$ when the plumber provides good service. Both probability distributions are computed according to Eq. (1). The incentive-compatible, collusion-resistant payments are the following: $\tau(0, 0) = \tau(0, 2) = 0$, $\tau(0, 1) = 12.37$, $\tau(0, 3) = \varepsilon$, $\tau(1, 0) = \varepsilon$, $\tau(1, 1) = \tau(1, 3) = 0$, and $\tau(1, 2) = 6.29$. ε can take any value greater than 0, and the guaranteed margin for truth-telling is $\delta = 1$.

5. NON-TRANSFERABLE UTILITIES, FULL COORDINATION

The next collusion scenario we are considering is when the N agents can use side-channels to coordinate their reporting strategies, but they cannot transfer payments from one another. Unlike the previous setting, here each of the N agents can have a different reporting strategy. The collusion strategy profile $a = (a_i)$, $i = 1, \ldots, N$ is no longer symmetric, and prescribes that agent i reports according to the strategy $a_i \in A$.

We distinguish between two cases, where the communication (and therefore the coordination on the reporting strategy profile) happens *before* or *after* the agents perceive the quality signals from the product they purchase. In both cases, however, we obtain negative results.

PROPOSITION 1. *When agents communicate and coordinate their reports* after *perceiving the quality signals, strict incentive-compatible payment mechanisms do not exist.*

PROOF. Consider two settings, that are identical except for the observation of agent i. In setting A, agent i observes $o_i = 0$, in setting B, agent i observes $o_i = 1$; in both A and B, the other agents observe n high quality signals. An incentive-compatible mechanism requires i to report 0 in setting A, and 1 in setting B. Assume all other agents report truthfully; during the communication phase (happening after signals have been perceived) agent i learns in both settings that the reference reports contain n positive reports. An incentive-compatible payment mechanism requires that:

- $\tau(0, n) > \tau(1, n)$ - honest reporting is strictly better for i in setting A ;
- $\tau(1, n) > \tau(0, n)$ - honest reporting is strictly better for i in setting B;

Clearly this is impossible. ■

The previous proposition formalizes the intuition that truth-telling may only be an ex-ante Nash equilibrium. The reference reports must be unknown to the agent in order to allow the design of incentive-compatible payments. When the communication takes place before the agents observe the signals, incentive-compatible payments do exist, but they always accept lying equilibria as well:

PROPOSITION 2. *When agents communicate and coordinate their reports* before *perceiving the quality signals, no payment mechanism has a unique honest reporting Nash equilibrium.*

PROOF. The proof shows that a full coalition can always find a profile of constant reporting strategies, $a = (a_i)$, $i = 1, \dots, N$, $a_i \in \{all0, all1\}$ that is a NEQ.

We define the family of reporting strategy profiles $a(n) = (a_i)$ where n out of N agents always report 1, and the other $N - n$ agents always report 0: i.e.,

$$\begin{aligned} &a_i = all1,\ \forall i \in S_1; \quad a_i = all0,\ \forall i \in S_0; \\ &|S_1| = n, \quad |S_2| = N - n; \\ &S_1 \cap S_0 = \varnothing; \quad S_1 \cup S_0 = \{1, 2, \dots, N\}; \end{aligned} \tag{9}$$

Assume that the payment mechanism defined by $\tau(\cdot, \cdot)$ accepts honest reporting as the unique NEQ. We have seen in Section 3 that the incentive-compatible constraints (4) imply the existence of $n_1 \neq n_2 \in \{0, 1, \dots, N-1\}$ such that $Pr[n_1|0] > Pr[n_1|1]$, $\tau(0, n_1) > \tau(1, n_1)$, $Pr[n_2|1] > Pr[n_2|0]$ and $\tau(1, n_2) > \tau(0, n_2)$.

With non-transferable utilities, the strategy profile $a(n_2 + 1)$ is not a NEQ if and only if one of the $n_2 + 1$ agents that should report 1 would rather report 0:

$$\tau(0, n_2) > \tau(1, n_2);$$

or one of the $N - n_2 - 1$ agents that should report 0 would rather report 1:

$$\tau(1, n_2 + 1) > \tau(0, n_2 + 1);$$

The first inequality cannot be true by the choice of n_2; therefore, it must be that $\tau(1, n_2 + 1) > \tau(0, n_2 + 1)$.

Similarly, $a(n_2 + 2)$ is not a NEQ iff either $\tau(0, n_2 + 1) > \tau(1, n_2 + 1)$ (impossible), or $\tau(1, n_2 + 2) > \tau(0, n_2 + 2)$. Continuing this argument we find that $\tau(1, N-1) > \tau(0, N-1)$ which makes $a(N)$ (i.e., all agents report 1) a Nash equilibrium. Hence the result of the proposition. ■

Proposition 2 holds regardless of the number of reports, N, available to the reputation mechanism. A full coalition can always find a lying reporting strategy profile that is a Nash equilibrium. By definition such a coalition is stable, i.e., no colluder has the incentives to deviate from the collusion strategy.

The obvious question is whether such coalitions are also *profitable*. Unless the collusion strategy brings every agent at least the payoff expected from the honest equilibrium, there may be reasons to believe that the coalition will never form. Profitable coalitions require lying Nash equilibria that pareto-dominate the honest one. A payment mechanism where such equilibria do not exist, is, in some sense, collusion-resistant.

Take for example the incentive-compatible payment scheme that solves LP 1, with the additional constraints that $n_1 \neq 0$ and $n_2 \neq N - 1$. A stable coalition can form on the strategy profile $a(n_2 + 1)$ (or $a(n_1)$), where $n_2 + 1$ (respectively n_1) agents report 1 and the others report 0, regardless of their observation. This equilibrium, however, does not pareto-dominate the truthful one: the agents that report 0 do not get any reward, whereas they do get rewarded in the honest equilibrium.

The payment mechanism can be further improved by setting $\tau(0, n_1 - 1) = \tau(1, n_2 + 1) = \varepsilon$, where ε is some small value. This modification eliminates the equilibria $a(n_2 + 1)$ and $a(n_1)$ and instead introduces the equilibria $a(n_2+2)$ and $a(n_1 - 1)$. Both these equilibria are extremely unattractive (some agents get paid ε, while others don't get paid at all) and are dominated by the honest one.

For any given strategy profile, a, either of the following linear constraints makes the payment mechanism resistant against a coalition on a:

$$\begin{aligned} &V(a_i, a_{-i}|o_i) < V(a_i^*, a_{-i}|o_i) \text{ for some } i, o_i \text{ and } a_i^*; \\ &V(a_i, a_{-i}|o_i) < V(h_i, h_{-i}|o_i) \text{ for some } i \text{ and } o_i; \end{aligned}$$

The first constraint ensures that a is not NEQ, the second that a does not pareto-dominate the honest equilibrium. Unfortunately, considering all strategy profiles is computationally infeasible. For this reason, we advocate an iterative solution, where the mechanism designer first solves LP 1 (with the additional constraints discussed in the previous paragraph), and then iteratively adds the constraints that eliminate lying pareto-optimal equilibria. This algorithm resembles the *incremental mechanism design* described by Conitzer and Sandholm [4] for social choice problems. As part of future research we plan to look for heuristics that help a designer select a small set of strategies that generate enough constraints to make honest reporting pareto optimal.

6. NON-TRANSFERABLE UTILITIES, PARTIAL COORDINATION

The setting described in Section 5 is very close to the worst-case collusion scenario that may be observed in online reputation mechanisms. The coalition comprises all reporters, and the coordination mechanisms are perfect.

In most practical applications, not all agents can collude. Some agents are altruistic in nature and report honestly for moral or social reasons. Other agents may not be contacted by a coalition. Social or legal condemnation of collusion may furthermore create prejudices that deter some agents from entering the coalition.

It is therefore reasonable to assume that a mechanism that prevents a fraction of the agents from colluding (while relying on the remaining fraction to report honestly) is good enough for most practical applications. The remaining question is whether the negative result of Proposition 3 still holds when some of the agents are unconditionally reporting the truth. As in the previous sections, agents may not transfer utilities from one another.

The existence of honest reports should help the reputation mechanism provide stronger truth-telling incentives. Indeed, *trusted reports* (reports generated by specialized, trusted reviewers that are hired to rate the product) help deter lying coalitions [13]. In this section we extend our previous work, by assuming that honest information comes from altruistic reporters, instead of being explicitly purchased. The main difference is that the reputation mechanism cannot identify the honest reports which get diluted in the group of all feedback reports.

When designing collusion-resistant, incentive-compatible payments we only consider the reporting strategy of a fraction containing k colluding agents. As the other $N - k$ agents report honestly, we can use stronger solution concepts than Nash equilibrium. When k is small enough, the honest reporting strategy (Nash equilibrium for the overall set of agents) may becomes a dominant strategy for the members of the coalition. Regardless of what the other $k-1$ agents report, truth-telling pays better than lying by some margin δ.

When n is the number of positive reports submitted by the $N - k$ honest reporters, and c is the number of positive reports submitted by the other $k-1$ colluders, the payments $\tau(\cdot, \cdot)$ satisfy the following constraints:

$$\begin{aligned}&\sum_{n=0}^{N-k} Pr[n|0]\big(\tau(0, n+c) - \tau(1, n+c)\big) \geq \delta;\\&\sum_{n=0}^{N-k} Pr[n|1]\big(\tau(1, n+c) - \tau(0, n+c)\big) \geq \delta;\end{aligned} \tag{10}$$

for all integers $c \in \{0, \dots, k-1\}$.

The objective function 5, and the set of constraints (10) form a linear optimization problem that defines the incentive-compatible payments that are resistant to a coalition of size k.

The remaining question is how large may the colluding fraction be, such that collusion-resistant, incentive-compatible mechanisms exist.

Proposition 3. *When more than k agents collude, with $2k > N$, no incentive-compatible payment mechanism can make truth-telling the dominant strategy for the colluders.*

Proof. The intuition behind the proof is the following: When $2k > N$, the $k - 1$ colluders submit at least as many reports as the remaining $N - k$ honest reporters. Therefore, any sequence of honest reports, can be 'corrected' by a carefully chosen sequence of colluding reports, such that lying is profitable.

Formally, let us take the subset $c = \{0, \dots, N - k\}$ (this subset exists because $N - k < k - 1$) from the system of inequalities defined by (10), and form the following optimization problem:

$$\begin{aligned}min \quad & W = Pr[1]\sum_{n=0}^{N-1} Pr[n|1]\tau(1,n) + Pr[0]\sum_{n=0}^{N-1} Pr[n|0]\tau(0,n);\\ s.t. \quad & \sum_{n=0}^{N-k} Pr[n|0]\big(\tau(0, n+c) - \tau(1, n+c)\big) \geq \delta;\\ & \sum_{n=0}^{N-k} Pr[n|1]\big(\tau(1, n+c) - \tau(0, n+c)\big) \geq \delta;\\ & \tau(0,n), \tau(1,n) \geq 0; \forall n = \{0, 1, \dots, N-1\};\end{aligned}$$

Let y_c^0 and y_c^1 be the dual variables corresponding to the constraints where the colluding agents report c positive signals, and the agent observes 0, respectively 1; One can easily verify that the dual problem accepts as solutions $y_c^1 = Pr[c|1] \cdot C$, $y_c^0 = Pr[c|0] \cdot C$, for all positive values C. The dual problem is therefore unbounded, which makes the primal infeasible. ■

The bound from Proposition 3 is also tight. Consider the example presented in the introduction, and assume the reputation mechanism has $N = 4$ reports. The following payments are resistant to the collusion of $k = 2$ agents: $\tau(0,0) = 1.575$, $\tau(0,1) = 3.575$, $\tau(0,2) = \tau(0,3) = 0$, $\tau(1,0) = \tau(1,1) = 0$, $\tau(1,2) = 2.203$, $\tau(1,3) = 0.943$. For example, if the client observes 1, reporting 1 is better than reporting 0 for any report of the other colluder:

$$\begin{aligned}&Pr[0|1]\tau(1,0) + Pr[1|1]\tau(1,1) + Pr[2|1]\tau(1,2) = 1.715;\\&Pr[0|1]\tau(0,0) + Pr[1|1]\tau(0,1) + Pr[2|1]\tau(0,2) = 0.715;\\&Pr[0|1]\tau(1,1) + Pr[1|1]\tau(1,2) + Pr[2|1]\tau(1,3) = 1.138;\\&Pr[0|1]\tau(0,1) + Pr[1|1]\tau(0,2) + Pr[2|1]\tau(0,3) = 0.138;\end{aligned}$$

where $Pr[0 \dots 2|1] = [0.0385, 0.1830, 0.7785]$ are the probabilities that 0, 1, or 2 out of 2 honest reports are positive, given that the client observed high quality.

6.1 The marginal cost of collusion resistance

Incentive-compatible payments that are resistant to coalitions of size k, must satisfy the constraints in (10). As k increases, the design problem becomes more constrained, and therefore, the budget required by the reputation mechanism is likely to grow. In this section we study the dependence of the expected budget on the size of the maximum tolerated coalition.

For a given context, the optimization problem that defines the payments $\tau_k(\cdot, \cdot)$ that are resistant to coalitions of size k is:

LP 2.

$$\begin{aligned}min \quad & Pr[1]\sum_{n=0}^{N-1} Pr[n|1]\tau_k(1,n) + Pr[0]\sum_{n=0}^{N-1} Pr[n|0]\tau_k(0,n);\\ s.t. \quad & \sum_{n=0}^{N-k} Pr[n|0]\big(\tau_k(0, n+c) - \tau_k(1, n+c)\big) \geq \delta;\\ & \sum_{n=0}^{N-k} Pr[n|1]\big(\tau_k(1, n+c) - \tau_k(0, n+c)\big) \geq \delta;\\ & \forall c \in \{0, \dots k-1\},\\ & \tau_k(0,n), \tau_k(1,n) \geq 0; \forall n = \{0, 1, \dots, N-1\};\end{aligned}$$

Table 1: Distribution of the maximum coalition bound. $\hat{k} = \lfloor N/2 \rfloor$ is the theoretical bound.

	Distribution of max coalition size (in %) over $[\hat{k}, \hat{k}-1, \ldots, 1]$
$N = 6, \hat{k} = 3$	[99.9, 0.08, 0.02]
$N = 11, \hat{k} = 5$	[99.42, 0.44, 0.1, 0.04, 0]
$N = 16, \hat{k} = 8$	[98.14, 0.68, 0.5, 0.36, 0.22, 0.06, 0.04, 0]
$N = 21, \hat{k} = 10$	[97.32, 0.82, 0.52, 0.44, 0.4, 0.3, 0.1, 0.06, 0.04, 0]
$N = 26, \hat{k} = 13$	[94.96, 1.48, 0.98, 0.62, 0.46, 0.44, 0.3, 0.28, 0.32, 0.06, 0.06, 0.04, 0]

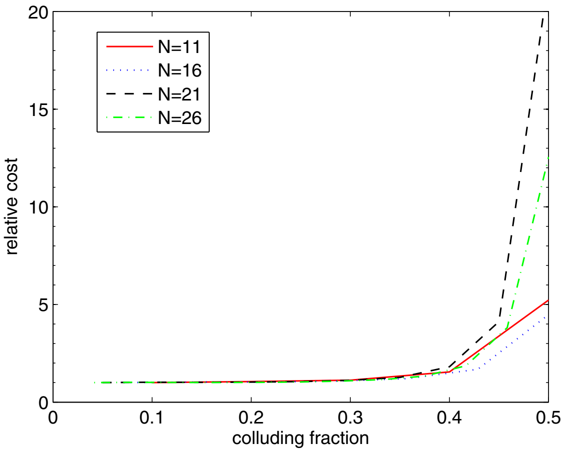

Figure 1: The relative cost of the mechanism as we increase the resistance to colluding fraction.

We numerically solved the optimal payment mechanism for 5000 problems, generated randomly in the following way:

- the set of possible types is randomly chosen between 2 and 20;
- for each type, θ, the probability, $p(\theta)$, that the buyers observe high quality is randomly chosen between 0 and 1;

We considered mechanisms for 6, 11, 16, 21 and 26 reports. As further evidence that the bound set by Proposition 3 is tight, Table 1 shows the distribution of the maximum collusion threshold among the problems we have solved. For example, when $N = 26$ reports, approximately 95% of the problems accept the theoretical bound described by Proposition 3.

Figure 1 plots the relative cost of the collusion-resistant mechanism (i.e., divided by the cost of the mechanism that is not collusion-resistant) as we increase the colluding fraction. It can be seen that the cost starts increasing exponentially when the payment mechanism must deter coalitions of more than one third of the agent population. This suggests that the practical bound on the coalition size should be around one third of the number N of reporters.

7. TRANSFERABLE UTILITIES, PARTIAL COORDINATION

As a last scenario we assume that colluding agents can redistribute the revenues among themselves. This will typically be the case when the same strategic agent controls a number of fake online identities (or *sybils* [1]). From the agent's perspective, the individual revenues obtained by each sybil is irrelevant; the objective of the agent is to maximize the cumulated revenue obtained by all sybils.

The fact that utilities are transferable, makes the problem of the mechanism designer significantly harder. In all previous scenarios, the constraints that made an incentive-compatible mechanism collusion-resistant ensured that lying coalitions are unstable: at least one of the colluders is better off by deviating from the colluding strategy. However, in this context the agents that suffer from following the colluding strategy may be rewarded by the others. The necessary (and sufficient) condition for collusion resistance requires that the cumulated revenue of the coalition is maximized when reporting the truth.

Another difference from the settings in Sections 5 and 6 is that colluders coordinate their reporting strategy *after* observing the quality signals. This assumption is supported by the interpretation that one strategic entity controls several fake online identities.

Concretely, we are looking for a payment mechanism with the following property: whenever k colluding agents observe c high quality signals, their cumulated revenue is maximized when reporting c positive reports. An underlying assumption is that non-colluders (the other $N-k$ agents) are reporting honestly. The revenue of the coalition that reports r (out of k) can be computed as follows. The r colluders that report positively are rewarded $\tau(1, r-1+n)$, while the $k-r$ colluders that report negatively are rewarded $\tau(0, r+n)$; n is the number of positive reports submitted by the (honest) non-colluders. The expected revenue of the coalition is therefore:

$$V(r|c) = \sum_{n=0}^{N-k} Pr[n|c]\Big(r \cdot \tau(1, r-1+n) + (k-r) \cdot \tau(0, r+n)\Big);$$

where $Pr[n|c]$ is the probability that n out of $N-k$ agents observe high quality signals, given that c out of k positive signals have already been observed.

Honest reporting is the best strategy for the coalition, when for all $c \in \{0, \ldots k\}$, $\arg\max_r V(r|c) = c$:

$$\sum_{n=0}^{N-k} Pr[n|c]\Big(c \cdot \tau(1, c-1+n) + (k-c) \cdot \tau(0, c+n) \\ - r \cdot \tau(1, r-1+n) - (k-r) \cdot \tau(0, r+n)\Big) \geq \delta; \tag{11}$$

The cheapest incentive-compatible, collusion-resistant payment mechanism minimizes the objective function (5) under the linear constraints (11):

LP 3.

$$\begin{aligned} &min \quad W = Pr[1]\sum_{n=0}^{N-1} Pr[n|1]\tau(1,n) + Pr[0]\sum_{n=0}^{N-1} Pr[n|0]\tau(0,n); \\ &s.t. \quad (11) \textit{ is true}, \forall c, r \in \{0, \ldots k\}, c \neq r \\ &\qquad \tau(0,n), \tau(1,n) \geq 0; \forall n = \{0, 1, \ldots, N-1\}; \end{aligned}$$

We used numerical simulations to evaluate (a) the maximum size of the tolerated coalitions, and (b) the marginal cost of increasing collusion resistance. As in Section 6, we generated 5000 random problems and computed the optimal payments for $N =$ 6,11,16,21 and 26 reports. For each case,

Table 2: Distribution of the maximum tolerable coalition size.

	Distribution of max coalition size
$N = 6$	$[5 : 100\%]$
$N = 11$	$[10 : 99.14\%, 9 : 0.36\%, 8 : 0.28\%, 7 : 0.14\%]$
$N = 16$	$[15 : 97.56\%, 14 : 0.44\%, 13 : 0.32\%, 12 : 0.52\%]$
$N = 21$	$[20 : 96.04\%, 19 : 0.52\%, 18 : 0.36\%, 17 : 0.42\%]$
$N = 26$	$[25 : 94.24\%, 24 : 0.66\%, 23 : 0.58\%, 22 : 0.46\%]$

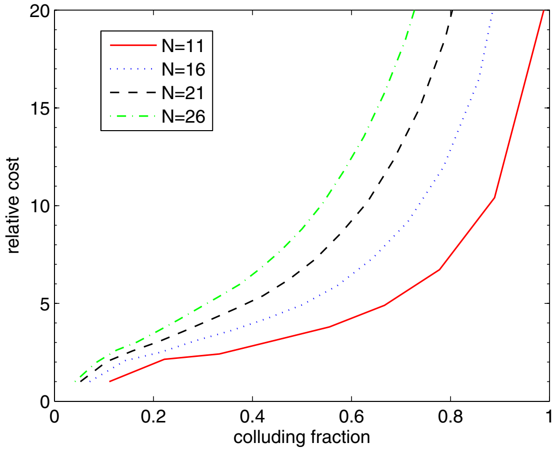

Figure 2: The relative cost of the mechanism as we increase the colluding fraction (setting with transferable utilities).

we gradually increased the coalition size (i.e., k) from 1 to $N - 1$.

Table 2 shows the distribution of the maximum coalition size that can be deterred by the payment mechanism. This threshold is most of the time equal to $N - 1$, meaning that one anonymous honest report is enough to design incentive-compatible, collusion-resistant payments. The result might be surprising when related to the bound of Proposition 3: an apparently more difficult collusion scenario allows the design of payments that are resistant to bigger coalition fractions. The explanation resides in the choice of the solution concept for the scenario in Section 6. There, honest reporting is the dominant strategy, so that each colluder reports the truth regardless of the reports of the other colluders. In the present scenario, on the other hand, individual colluders are allowed to lie, if the coalition as a whole reports the truth. The constraints of LP 3 are therefore feasible for higher coalition fractions.

The marginal cost of collusion resistance is however higher than in Section 6. Figure 2 plots the relative cost of the collusion-resistant mechanism (i.e., divided by the cost of the mechanism that is not collusion-resistant) as we increase the tolerated coalition fraction. The cost grows linearly for coalitions that span up to one half of the population; for larger coalitions, the cost grows exponentially. Nonetheless, by comparing Figures 2 and 1, we see that for the same coalition size, the collusion-resistant payments are cheaper if we assume a setting with non-transferable utilities.

One last thing we would like to point out is that realistic collusion scenarios are most likely a combination between the settings presented in Sections 6 and 7: several strategic agents, each controlling several fake identities, try to manipulate the reporting mechanism. It is encouraging to see that separately, in each scenario we can achieve relatively high collusion resistance at acceptable costs. As future work, we plan to use a combination of the two techniques to make the mechanisms even better.

8. CONCLUSION AND FUTURE WORK

As feedback forums and reputation mechanisms become increasingly important sources of information, explicit measures must guarantee that honest reporting is in the best interest of the participants. Previous work shows that it is possible to construct payment mechanisms that reward honest reports higher (in expectation) than false ones. Truthtelling thus becomes a Nash equilibrium.

Unfortunately, such mechanisms also have other equilibria where reporters lie. This creates collusion opportunities, since several agents can coordinate their lies in order to improve their revenues. In this paper we addressed the design of incentive-compatible payments that are also resistant to collusion. For each of the four considered collusion scenario, we defined a linear optimization problem that allows the automated design of the cheapest payment mechanism.

In Section 4 we showed that incentive-compatible payments can be efficiently constructed such that honest reporting is the unique pure strategy symmetric equilibrium. The results can be easily extended so that honest reporting becomes the pareto-optimal equilibrium. However, we only treated pure strategies. Preventing *mixed*-strategy symmetric equilibria from becoming NEQ requires non-linear constraints, that make the design problem computationally difficult. Since reputation mechanisms need to compute such payments for every context, we believe that the design problem should be kept simple, on the expense of precision. An interesting question, therefore, is if we can find simple (e.g., linear constraints), heuristic extensions to the linear program of Section 4 that (a) make unlikely the existence of mixed strategy symmetric equilibria, or (b) make unlikely the existence of mixed strategy symmetric equilibria that pareto-dominate the honest equilibrium.

For the scenario in Section 5, we proved that any incentive-compatible payment mechanism also accepts lying equilibria, and suggested an iterative approach for finding the payments where truth-telling is not dominated by any of the lying equilibria. As future work, we plan to describe practical algorithms that guide mechanism designers in choosing the best direction for incrementally improving the mechanism.

In Section 6 we took advantage of the assumption that some agents will inherently report the truth, and constructed an incentive-compatible mechanism with much stronger collusion guarantees. When less than half of the population colludes, it is theoretically possible to construct payments that make honest reporting the dominant strategy for the colluders. Numerical simulations show, however, that the practical bound is closer to one third of the population: preventing coalition fractions greater than one third requires exponentially higher budget. An interesting open question is if we can increase the theoretical (and practical) bound by requiring honest reporting to be the unique (or pareto-optimal) Nash equilibrium. The constraints that define the corresponding mechanisms lead to non-linear optimization problems that do not scale well. We plan to use approximations and heuristic methods to further investigate this question.

Finally, Section 7 described incentive-compatible payments that are resistant to *sybil* attacks: i.e., the same strategic agents creates several fake identities in order to manipulate the payment mechanism. The designer can ensure that the set of reports submitted by the coalition reflects the aggregated experience of the coalitions. Individual colluders do not necessarily report the truth, but overall, the reputation mechanism obtains correct information.

9. REFERENCES

[1] A. Cheng and E. Friedman. Sybilproof reputation mechanisms. In *Proceeding of the Workshop on Economics of Peer-to-Peer Systems (P2PECON)*, pages 128–132, 2005.

[2] R. T. Clemen. Incentive contracts and strictly proper scoring rules. *Test*, 11:167–189, 2002.

[3] V. Conitzer and T. Sandholm. Complexity of mechanism design. In *Proceedings of the Uncertainty in Artificial Intelligence Conference (UAI)*, 2002.

[4] V. Conitzer and T. Sandholm. Incremental Mechanism Design. In *Proceedings of the IJCAI*, 2007.

[5] J. Crémer and R. P. McLean. Optimal Selling Strategies under Uncertainty for a Discriminating Monopolist When Demands Are Interdependent. *Econometrica*, 53(2):345–61, 1985.

[6] R. K. Dash, N. R. Jennings, and D. C. Parkes. Computational-mechanism design: A call to arms. *IEEE Intelligent Systems*, pages 40–47, November 2003. Special Issue on Agents and Markets.

[7] C. d'Aspremont and L.-A. Grard-Varet. Incentives and Incomplete Information. *Journal of Public Economics*, 11:25–45, 1979.

[8] A. Harmon. Amazon Glitch Unmasks War of Reviewers. *The New York Times*, February 14, 2004.

[9] B. Holmström. Moral Hazard in Teams. *Bell Journall of Economics*, 13:324–340, 1982.

[10] M. Jackson. A crash course in implementation theory. *Social Choice and Welfare*, 18(4):655–708, 2001.

[11] M. Jackson. *Encyclopedia of Life Support Systems*, chapter Mechanism Theory. EOLSS Publishers, Oxford UK, 2003.

[12] S. Johnson, J. Pratt, and R. Zeckhauser. Efficiency Despite Mutually Payoff-Relevant Private Information: The Finite Case. *Econometrica*, 58:873–900, 1990.

[13] R. Jurca and B. Faltings. Enforcing Truthful Strategies in Incentive Compatible Reputation Mechanisms. In *Internet and Network Economics*, volume 3828 of *LNCS*, pages 268 – 277. 2005.

[14] R. Jurca and B. Faltings. Minimum payments that reward honest reputation feedback. In *Proceedings of the ACM Conference on Electronic Commerce*, Ann Arbor, Michigan, USA, June 11-15 2006.

[15] M. Kandori and H. Matsushima. Private observation, communication and collusion. *Econometrica*, 66(3):627–652, 1998.

[16] S. Li and K. Balachandran. Collusion proof transfer payment schemes with multiple agents. *Review of Quantitative Finance and Accounting*, 15:217–233, 2000.

[17] C. Ma. Unique implementation of incentive contracts with many agents. *Review of Economic Studies*, pages 555–572, 1988.

[18] N. Miller, P. Resnick, and R. Zeckhauser. Eliciting Informative Feedback: The Peer-Prediction Method. *Management Science*, 51:1359 –1373, 2005.

[19] A. Parasuraman, V. Zeithaml, and L. Berry. A Conceptual Model of Service Quality and Its Implications for Future Research. *Journal of Marketing*, 49:41–50, 1985.

[20] T. Sandholm. Automated mechanism design: A New Application Area for Search Algorithms. In *Proceedings of the International Conference on Principles and Practice of Constraint Programming*, 2003.

[21] E. White. Chatting a Singer Up the Pop Charts. *The Wall Street Journal*, October 15, 1999.

APPENDIX

A. ANALYTICAL SOLUTION FOR INCENTIVE-COMPATIBLE PAYMENT MECHANISMS

For solving LP 1, let us write the corresponding dual problem:

$$\begin{array}{ll} max & \delta y_0 + \delta y_1; \\ s.t. & Pr[n|0]y_0 - Pr[n|1]y_1 \leq Pr[0]Pr[n|0] \\ & Pr[n|1]y_1 - Pr[n|0]y_0 \leq Pr[1]Pr[n|1] \\ & \forall n \in \{0,\dots,N-1\}; \end{array}$$

where y_0 (respectively y_1) is the dual variable corresponding to the constraint where the agent observes 0 (respectively 1). By dividing the first set of constraints with $Pr[n|0]$ and the second set of constraints with $Pr[n|1]$, we have:

$$y_0 - y_1 Pr[n|1]/Pr[n|0] \leq Pr[0], \forall n \in \{0,\dots,N-1\};$$
$$y_1 - y_0 Pr[n|0]/Pr[n|1] \leq Pr[1], \forall n \in \{0,\dots,N-1\};$$

Clearly, among the $2(N-1)$ constraints of the dual problem, only two are active, corresponding to: $n_1 = \arg\min_n \frac{Pr[n|1]}{Pr[n|0]}$, and $n_2 = \arg\min_n \frac{Pr[n|0]}{Pr[n|1]}$. It follows that only two of the variables of LP 1 (i.e., $\tau(0,n_1)$ and $\tau(1,n_2)$) have positive values. These values can be computed by solving the system of linear equations:

$$Pr[n_1|0]\tau(0,n_1) - Pr[n_2|0]\tau(1,n_2) = \delta;$$
$$-Pr[n_1|1]\tau(0,n_1) + Pr[n_2|1]\tau(1,n_2) = \delta;$$